The Rich Transcription 2009 Speech-To-Text (STT) and Speaker Attributed STT (SASTT) Results

2009 Rich Transcription Evaluation Workshop

May 28-29, 2009

Florida Institute of Technology

Melbourne, FL

Jérôme Ajot & Jonathan Fiscus

http://itl.nist.gov/iad/mig/tests/rt/2009/

NIST
National Institute of Standards and Technology

Speech-To-Text (STT)

- Task:
  - Transcribe the spoken words
- Domain:
  - Conference Room (confmtg)
- Primary input condition:
  - Multiple Distant Microphones (MDM)
- Participating sites:
  - AMI, FIT, SRI/ICSI

STT Evaluation Protocol

- Step 1: Transcript normalization
  - Motivation: Allow acceptable alternative transcripts
    - Differentiating ***gonna*** from ***going to*** is sometimes difficult
  - Implementation: Text filtering rules applied to both the reference and system transcript

- Step 2: Overlapping Speech Text Alignment
  - Motivation: Identify and classify errors by finding an optimal one-to-one mapping of reference to system words

- Step 3: Error computation
  - Primary Metric: Word Error Rate (WER):

$$100 \cdot \frac{N_{Substitutions} + N_{Insertions} + N_{Deletions}}{N_{referenceWords}}$$

  - 0% is best possible score, more than 100% possible

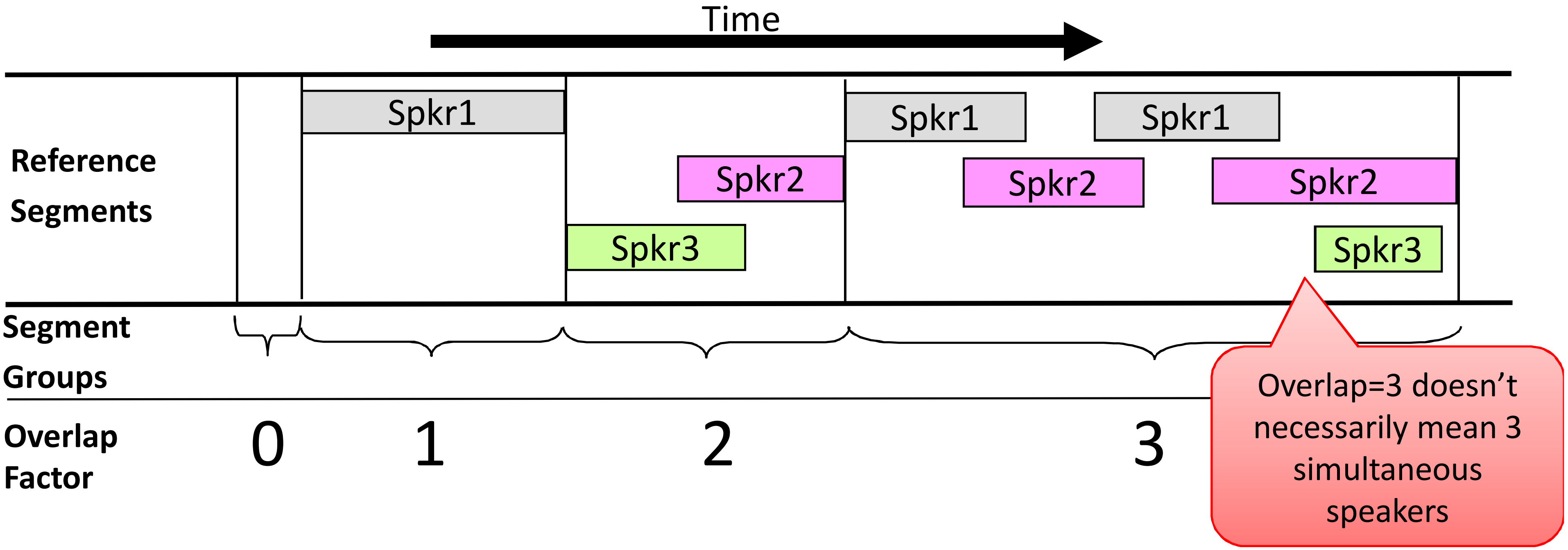
National Institute of Standards and Technology

Overlapping Speech Text Alignments

- Solution: Multi-dimensional text alignments produce the 1:1 mapping
  - Each speaker (reference and system) is a dimension in a Levenshtein Edit Distance matrix
  - Alignment engine implemented within ASCLITE
- Challenge: Computational complexity limits
  - Search space limited by applying heuristics
    - Pre-segmenting the reference transcript into “Segment Groups”
    - Heuristic pruning, application constraints, and memory compression
- Net Effect:
  - More evaluable data
  - Faster scoring time

Segment Groups

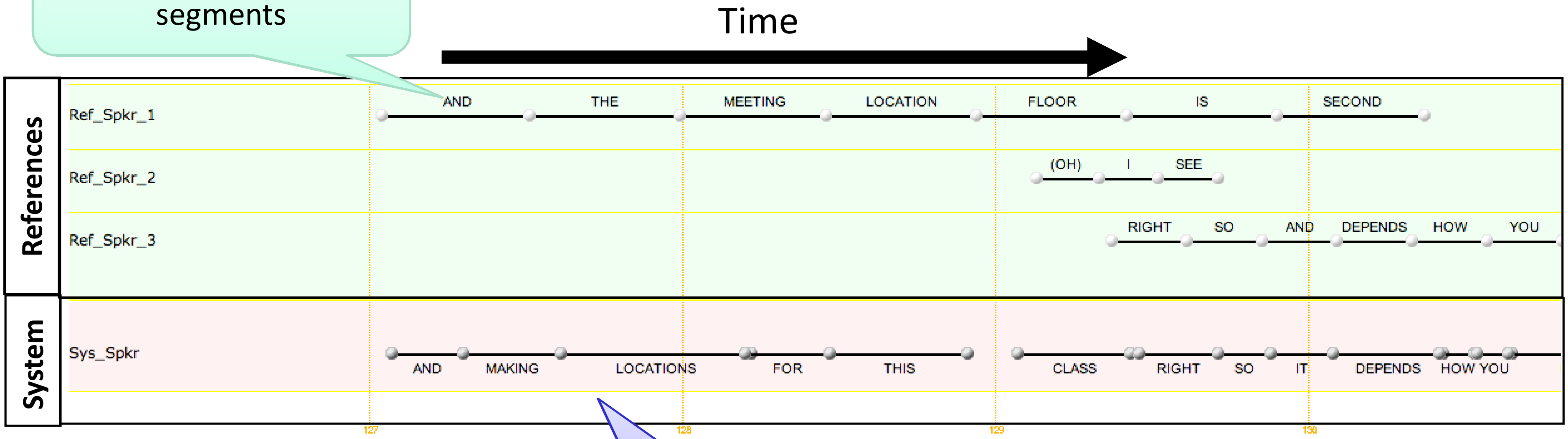
Divide the reference transcript segments into independent units based on segment times



- Smaller overlap factor → faster alignment times
- Overlap factors used for conditional scoring

Multi-Dimensional Alignment Visualization for STT

Interpolated word times within reference segments

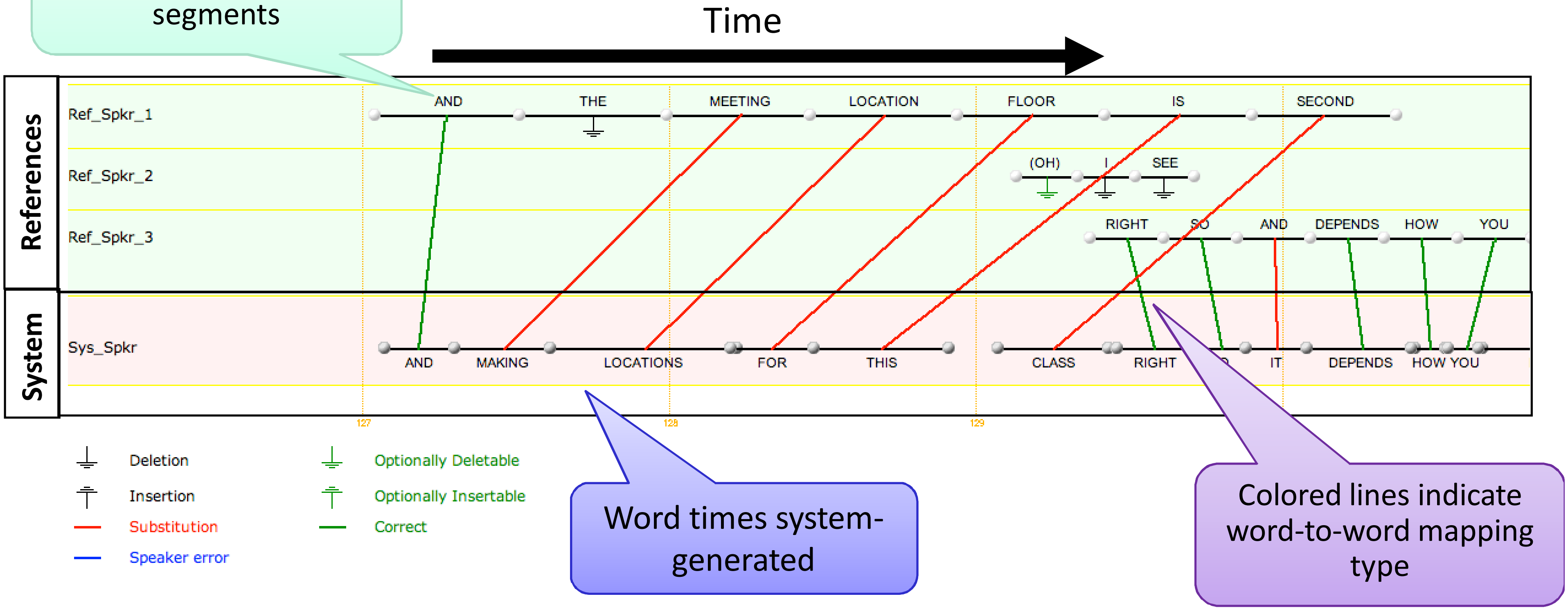


Word times system-generated

NIST
National Institute of Standards and Technology

Multi-Dimensional Alignment Visualization for STT

Interpolated word times within reference segments



Word times system-generated

Colored lines indicate word-to-word mapping type

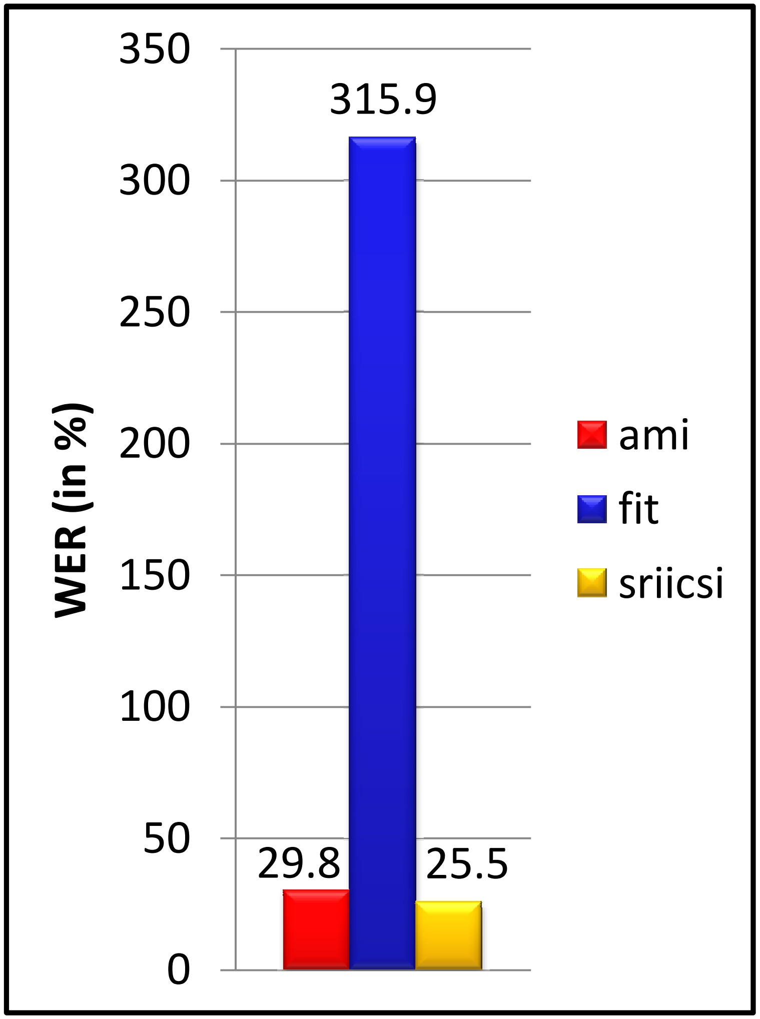
4 Dimensional Alignment labeled as Overlap = 3

0.12 MB to align

NIST National Institute of Standards and Technology

STT Primary System Results

IHM Condition

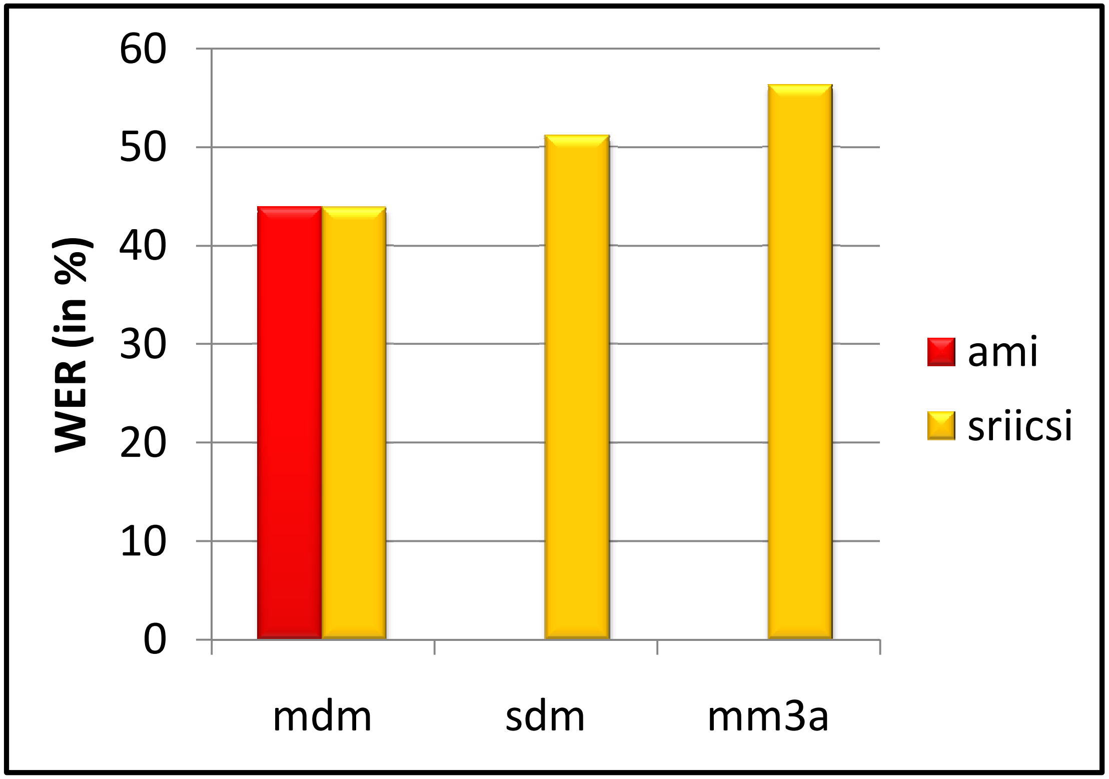


- 3 STT – IHM submission
- FIT is a first time participant

NIST
National Institute of Standards and Technology

STT Primary System Results

Distant Microphone (Overlap ≤ 4)

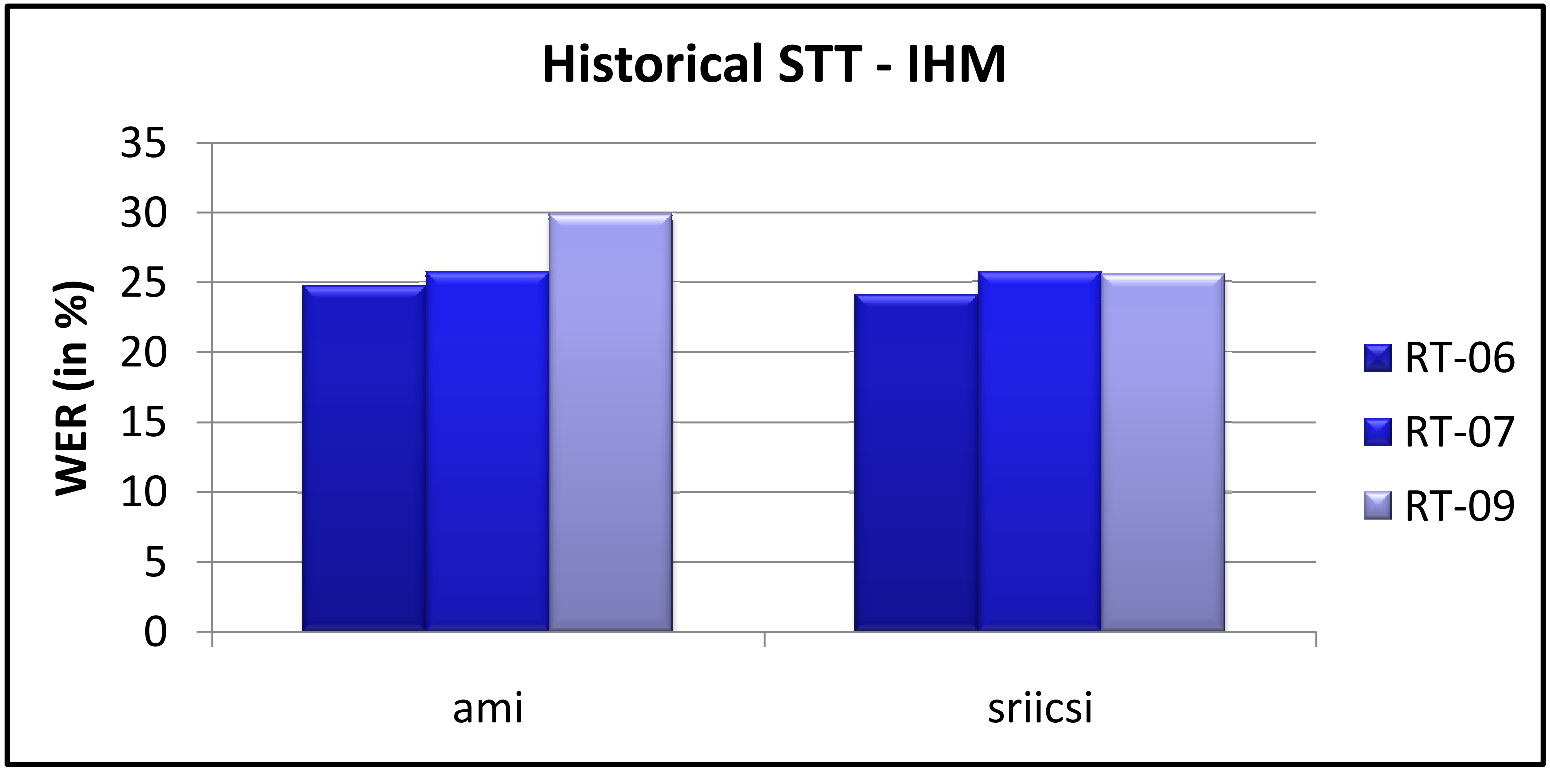


- Distant microphone conditions increase the difficulty
- SRIICSI is able to make use of distant microphones

NIST
National Institute of Standards and Technology

Historical STT Performance

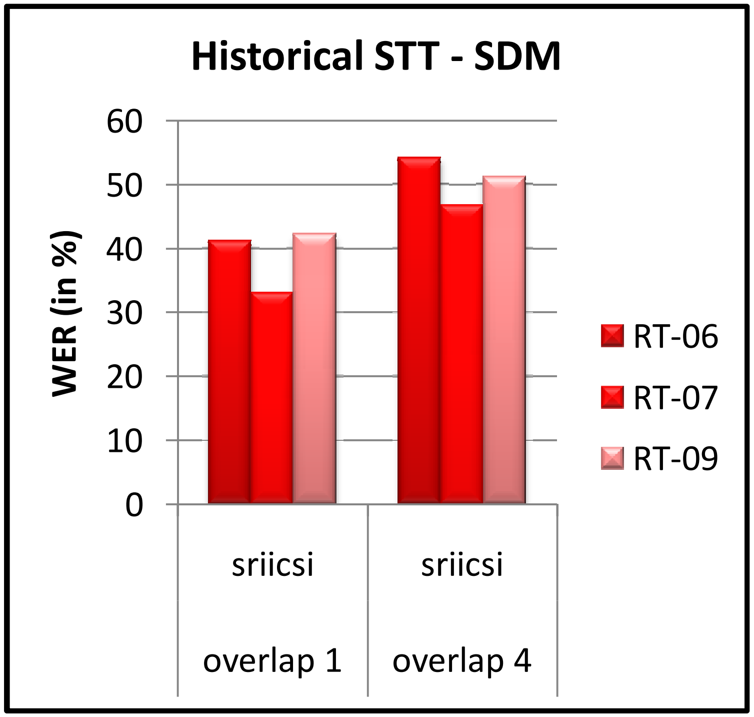
IHM Condition

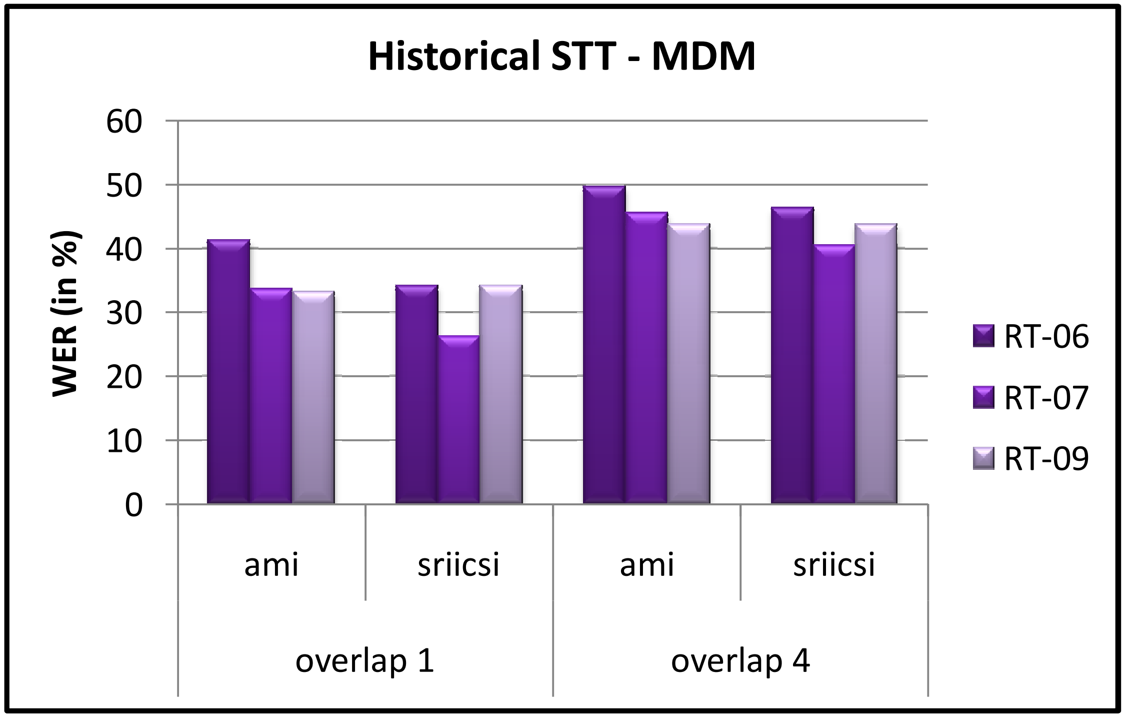


- IHM condition was challenging for AMI
- SRIICSI has a stable performance over the last 3 evaluations

Historical STT Performance

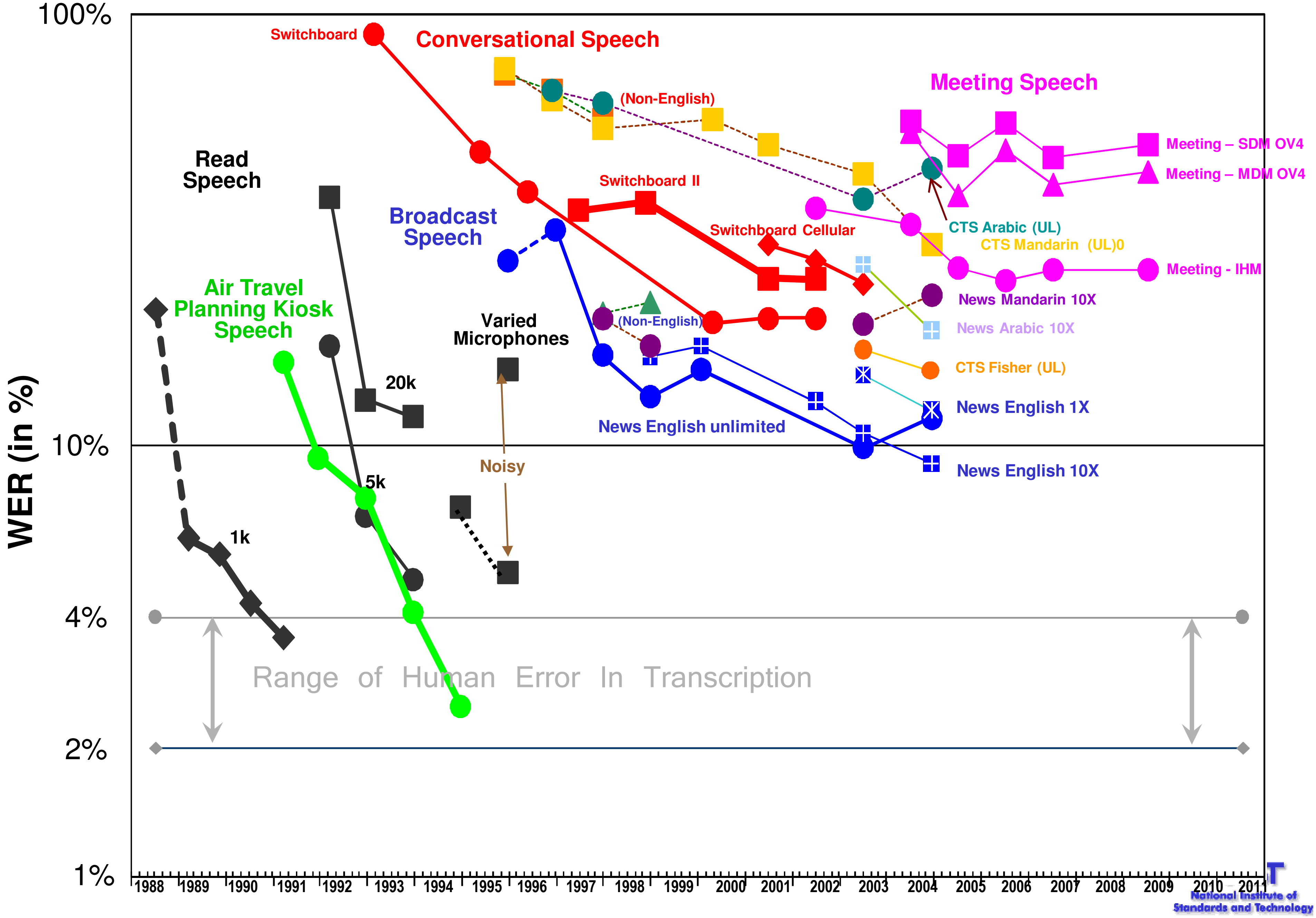
Distant Microphones





- AMI progressed over the last 3 evaluations for MDM
- Results are inconclusive for SRIICSI

NIST STT Benchmark Test History – May. ’09



National Institute of Standards and Technology

Speaker Attributed STT (SASTT)

- Task:
  - Transcribe the spoken words and associate them with a speaker
  - Merge of STT and Speaker Diarization systems
- Domain:
  - Conference Room (confmtg)
- Primary input condition:
  - Multiple Distant Mics (MDM)
- Participating sites:
  - AMI, SRI/ICSI

SASTT Evaluation Protocol

- Step 1: Transcript normalization
  - Identical to STT

- Step 2: Speaker Alignment
  - Define what is “correct” speaker
  - A one-to-one mapping between reference and system speakers
  - Same time-time based scoring method as used for the Speaker Diarization Task (SPKR)
    - Except system segments derived from recognized word locations

- Step 3: Text Alignment
  - A one-to-one mapping is found between the reference and system transcripts
  - Changes to mapping requirements
    - Correct: matching words and mapped reference/system speaker
    - **Speaker Substitution**: correct words and non-mapped reference/system speakers
    - Substitution: non-matching texts

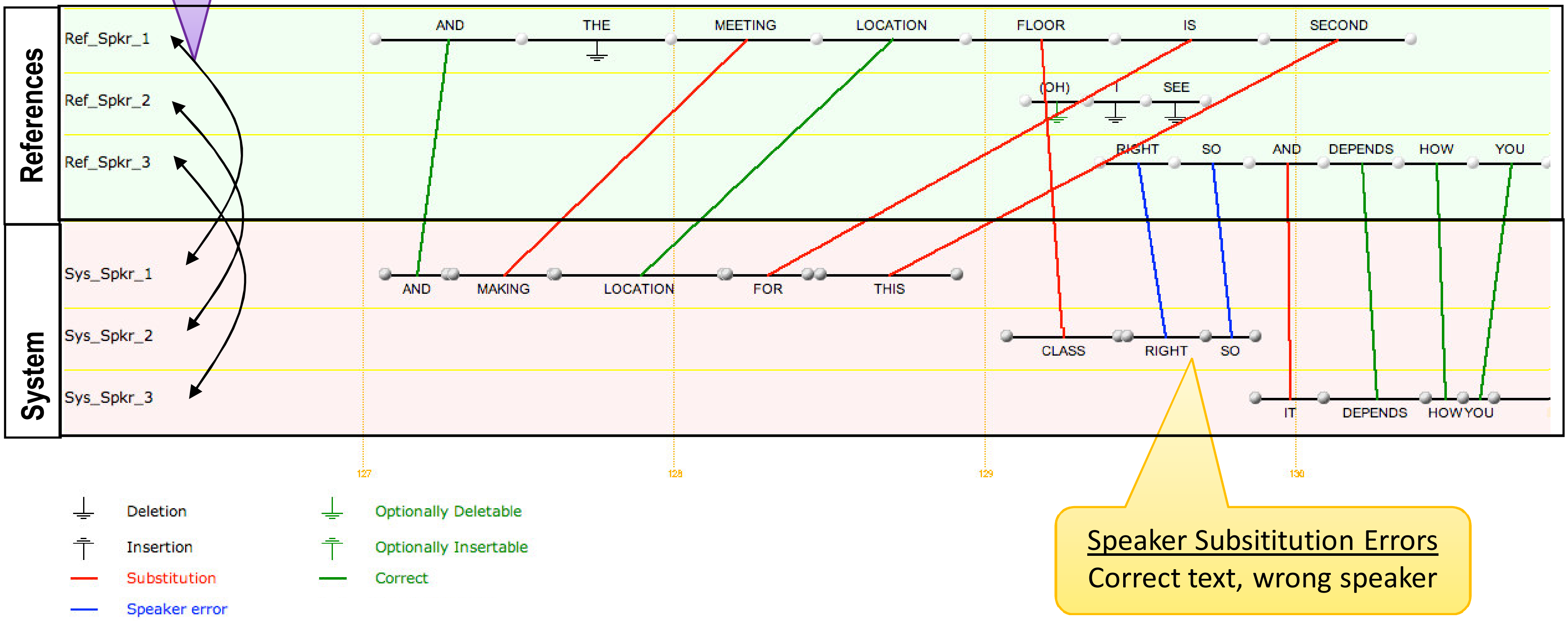
- Step 4: Error computation
  - Primary Metric: Speaker Attributed Word Error Rate (SWER):

$$100 \cdot \frac{N_{Substitutions} + N_{Insertions} + N_{Deletions} + N_{Spea\,ker\,Substitution}}{N_{reference\,Words}}$$

  - 0% is best possible score, more than 100% possible

Multi-Dimensional Alignment Visualization for SASTT

Mapped Speakers



Speaker Subsititution Errors

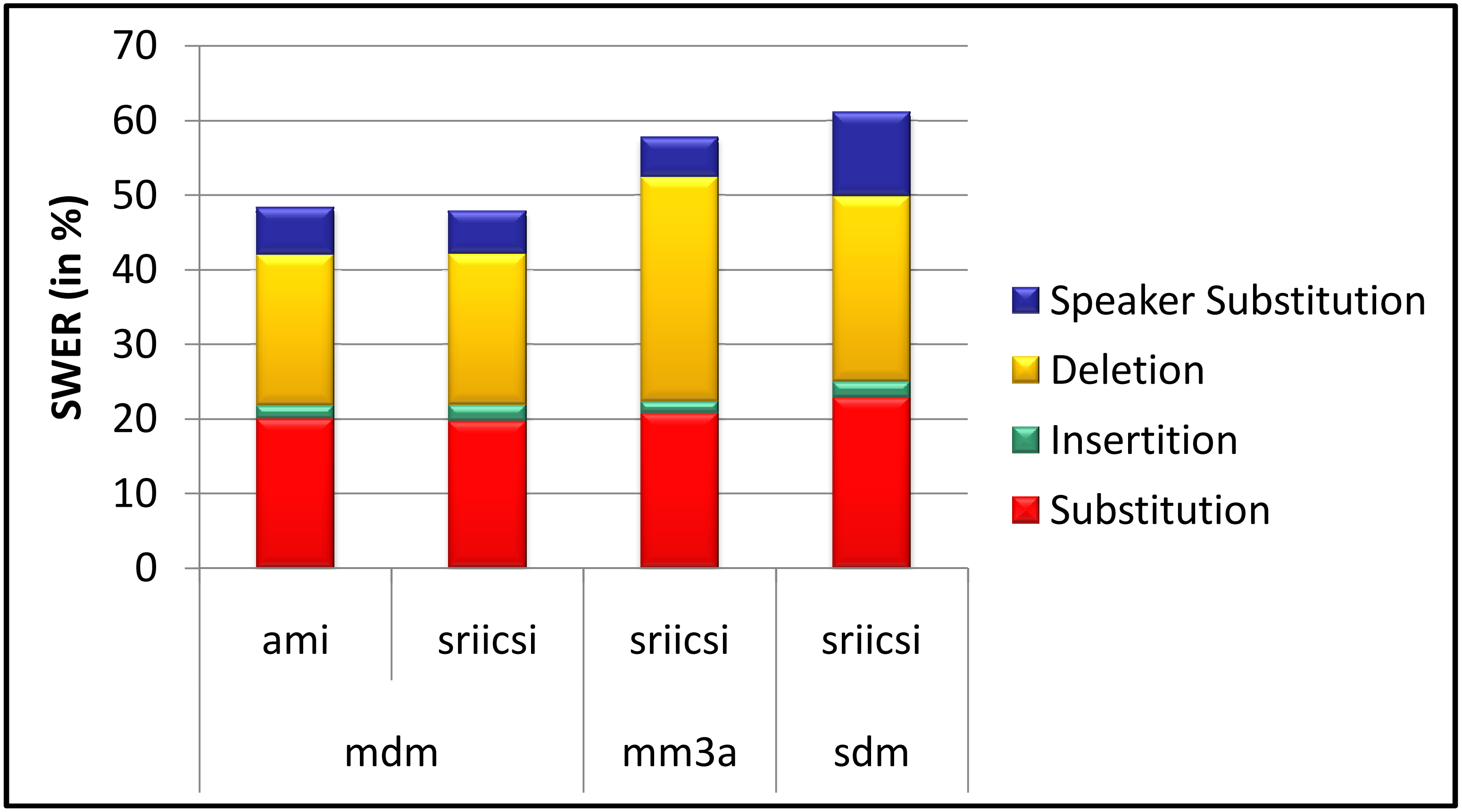
Correct text, wrong speaker

6 Dimensional Alignment labeled as Overlap = 3

2.12 MB to align → 18 times bigger than STT

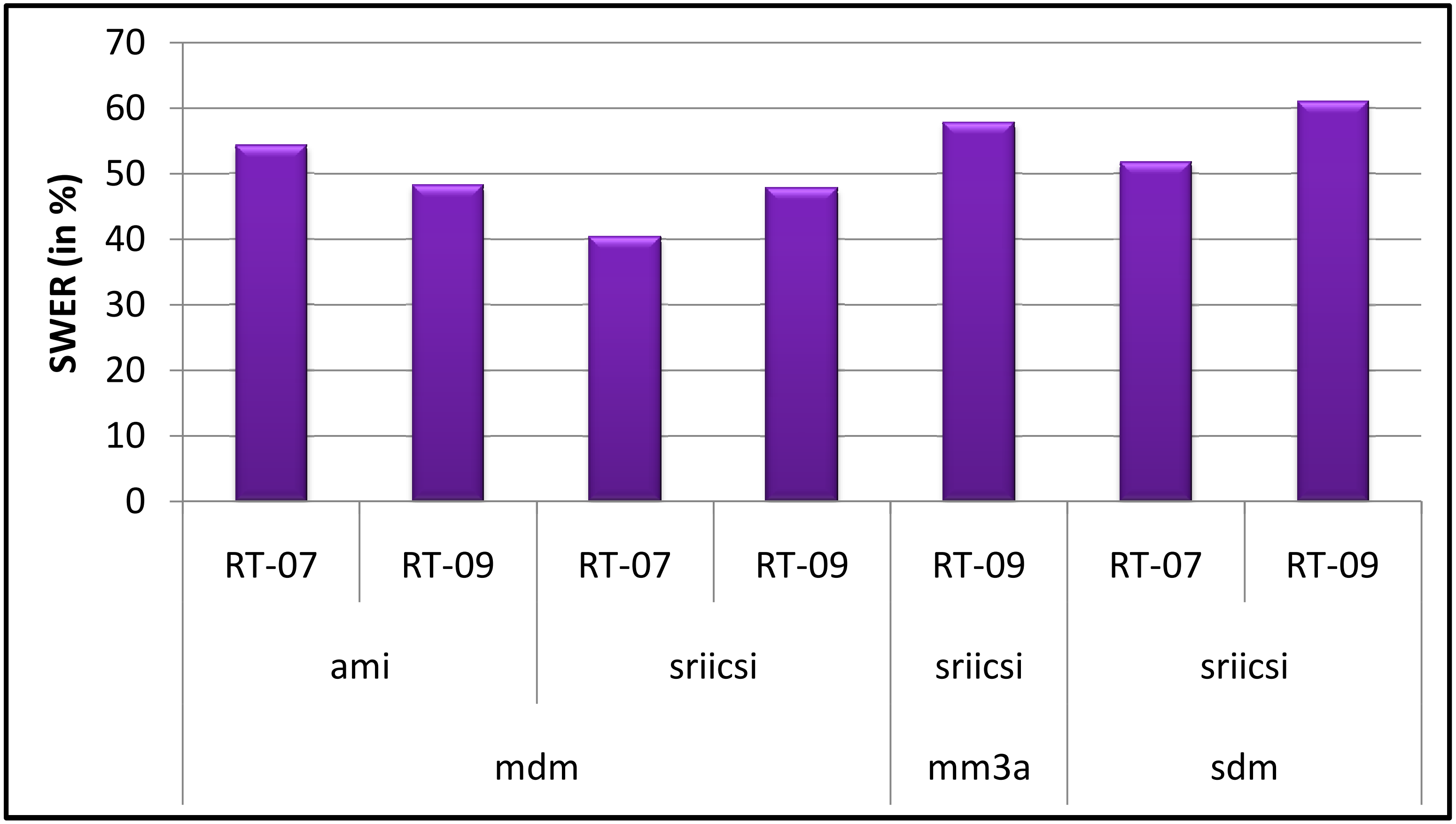
NIST
National Institute of Standards and Technology

SASTT Results
(≤ 3 speakers)



- As for STT, distant microphones are challenging conditions

SASTT Results (≤ 3 speakers)

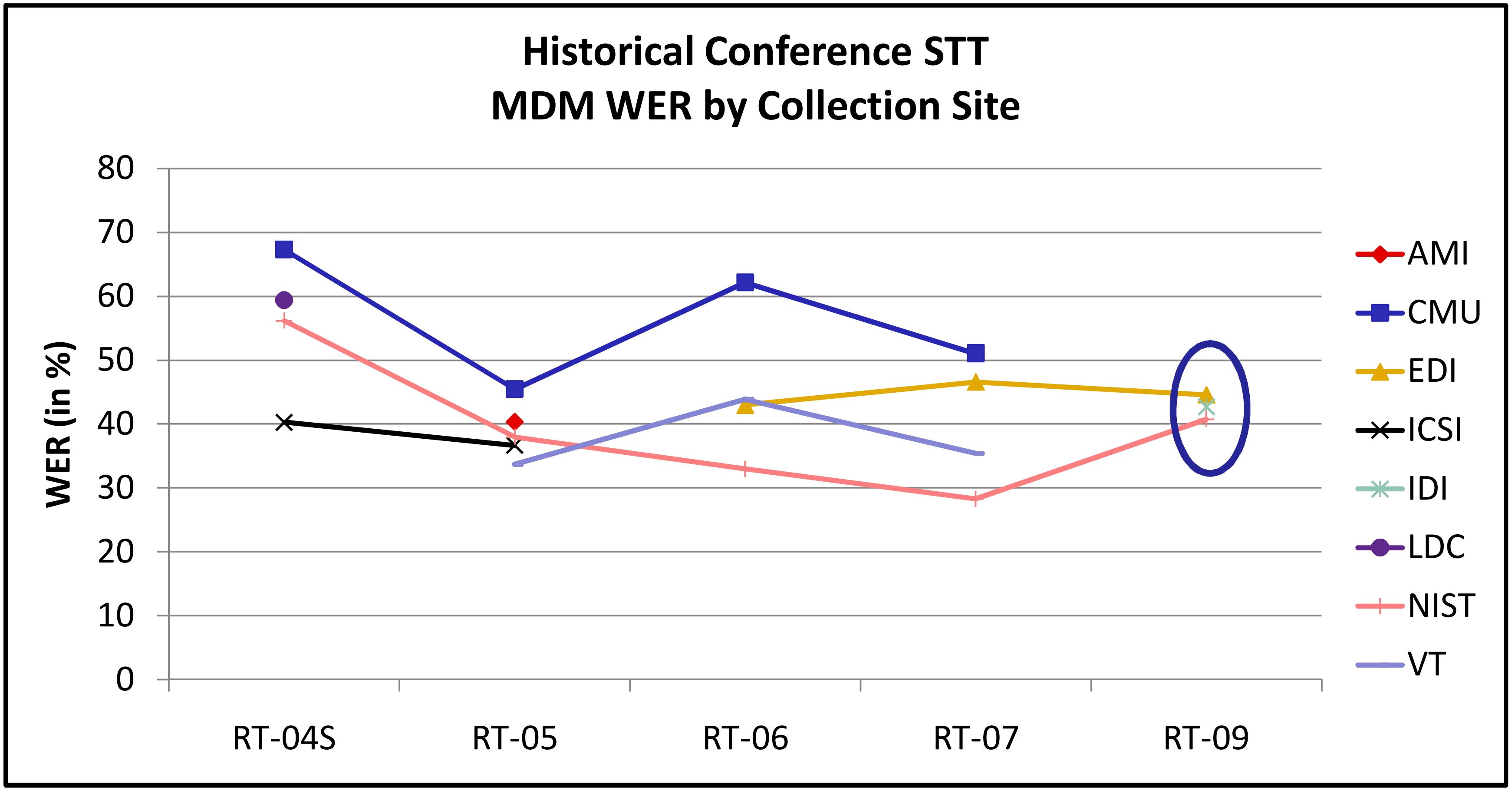


- Compared to last evaluation AMI progresses in the MDM condition
- But the test set was still chalenging

NIST National Institute of Standards and Technology

Test Sets

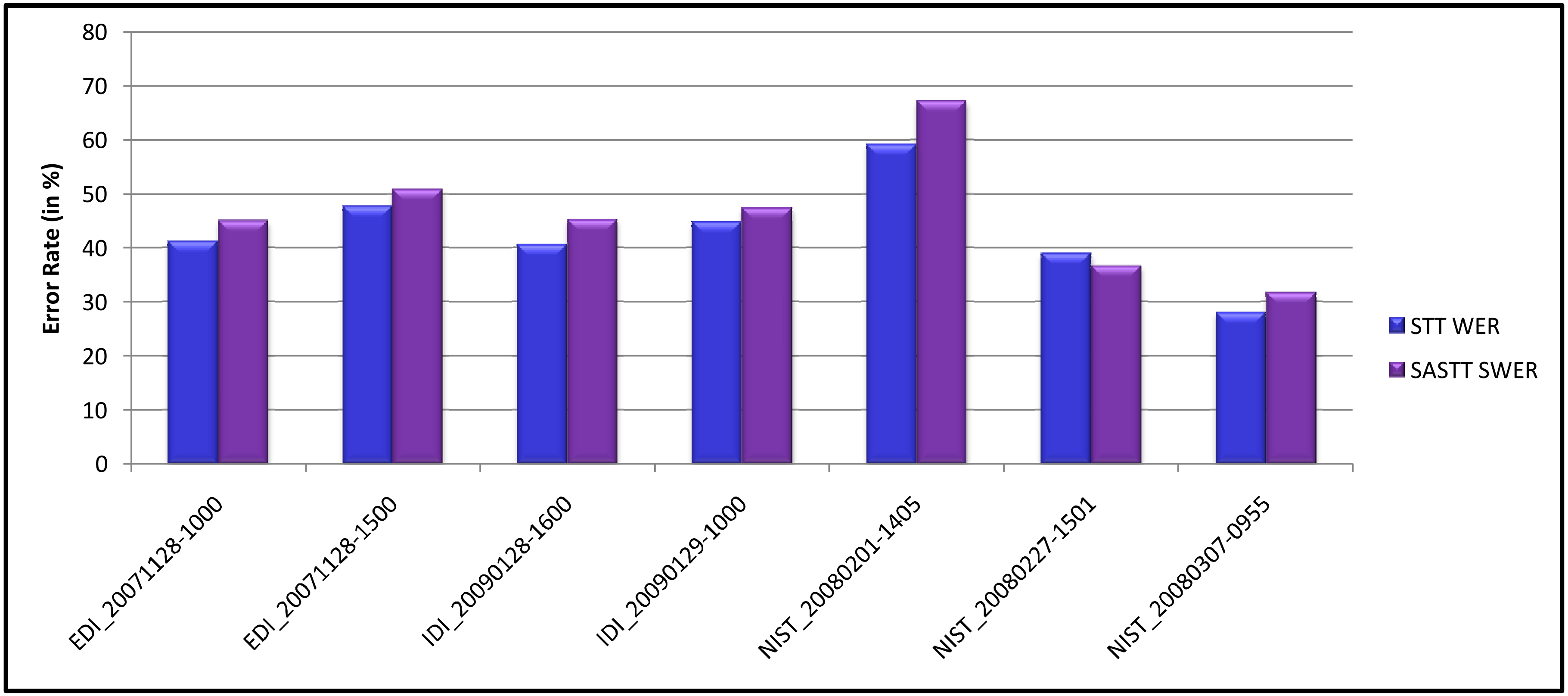
Collection Sites



- Little difference this year for STT – MDM by collection site

Test Sets

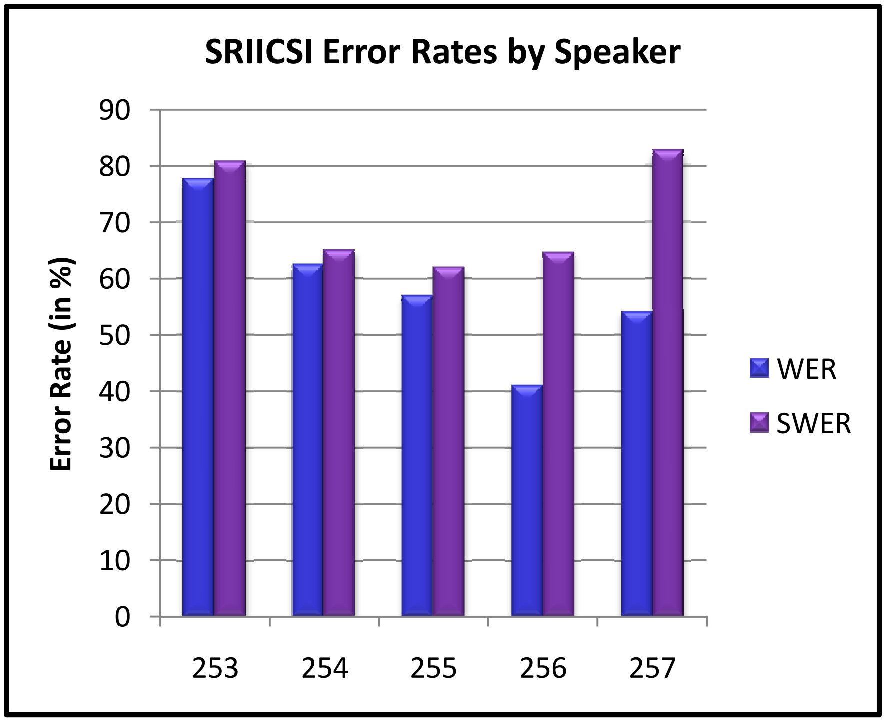
Meetings Variability



- Diversity in the meeting dialect
  - EDI and IDI meetings have only non-native American speakers
  - NIST meetings have only native American speakers
- Variability in the NIST meeting

Test Sets

NIST_20080201-1405



- High overlap factor meeting
- All speakers have high deletion rate: 25-60% *(average: 20%)*
- Speaker 256 and 257 have a high rate of Speaker Substitution Error: 23-27% *(average: 5%)*

NIST
National Institute of Standards and Technology

Conclusions

- RT-09 Results
  - No noticeable improvements
- Challenging test sets
- Future evaluations Data Set
  - More diverse test set
    - Small segments
    - More meetings
  - Progress test set
    - Sequestering data
- Focus on core technology challenges
  - Overlapping speech
  - Distant microphones